Why ability point estimates can be pointless: a primer on using skill measures from large-scale assessments in secondary analyses

Lechner, Clemens; Bhaktha, Nivedita; Groskurth, Katharina; Bluemke, Matthias

Measurement Instruments for the Social Sciences

ADVANCES IN METHODOLOGY Open Access

Why ability point estimates can be pointless: a primer on using skill measures from large-scale assessments in secondary analyses

Clemens M. Lechner*, Nivedita Bhaktha, Katharina Groskurth and Matthias Bluemke

Abstract

Measures of cognitive or socio-emotional skills from large-scale assessments surveys (LSAS) are often based on advanced statistical models and scoring techniques unfamiliar to applied researchers. Consequently, applied researchers working with data from LSAS may be uncertain about the assumptions and computational details of these statistical models and scoring techniques and about how to best incorporate the resulting skill measures in secondary analyses. The present paper is intended as a primer for applied researchers. After a brief introduction to the key properties of skill assessments, we give an overview over the three principal methods with which secondary analysts can incorporate skill measures from LSAS in their analyses: (1) as test scores (i.e., point estimates of individual ability), (2) through structural equation modeling (SEM), and (3) in the form of plausible values (PVs). We discuss the advantages and disadvantages of each method based on three criteria: *fallibility* (i.e., control for measurement error and unbiasedness), *usability* (i.e., ease of use in secondary analyses), and *immutability* (i.e., consistency of test scores, PVs, or measurement model parameters across different analyses and analysts). We show that although none of the methods are optimal under all criteria, methods that result in a single point estimate of each respondent's ability (i.e., all types of "test scores") are rarely optimal for research purposes. Instead, approaches that avoid or correct for measurement error—especially PV methodology—stand out as the method of choice. We conclude with practical recommendations for secondary analysts and data-producing organizations.

Keywords: Large-scale assessments, Measurement error, Test scores, Plausible values

Introduction

In the last two decades, large-scale assessments surveys (LSAS) have expanded considerably in number and scope. National and international LSAS, such as PISA, TIMMS, PIAAC, NEPS, or NAEP, now provide a wealth of data on cognitive and socio-emotional (or "non-cognitive") skills[1] of children, youth, and adults. This increasing data availability has led to a veritable surge in investigations in economics, psychology, and sociology on issues such as skill formation, inequality in skills, or labor market returns to skills.

As the methodological sophistication of LSAS has evolved, the gap between expert psychometricians[2] who curate the assessments and applied researchers who use these data as secondary analysts has widened. LSAS often apply advanced statistical models and scoring techniques with which few applied researchers are familiar (e.g., Jacob & Rothstein, 2016; Jerrim, Lopez-Agudo,

[1]In this paper, we use the terms "ability,", "skills," and "proficiency" interchangeably. Although there are some differences across (sub-)disciplines in the way these terms are used, these subtle differences are of no import for our present intent.

[2]Psychometrics is the field of study concerned with psychological measurement.

* Correspondence: clemens.lechner@gesis.org
Department of Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, PO Box 2 21 55, 68072 Mannheim, Germany

Marcenaro-Gutierrez, & Shure, 2017). Consequently, there is uncertainty among applied researchers about the statistical assumptions and computational details behind these different models and scoring techniques, their respective pros and cons, and how to best incorporate the skill measures that result from them in secondary analyses. Moreover, secondary analysts often use the best available methods in less-than-optimal ways (e.g., Braun & von Davier, 2017). Thus, there is the risk of a growing disconnect between best practices in the use of data from LSAS and actual practices in applied research. Less-than-optimal practices may result in faulty analyses and erroneous substantive conclusions.

Against this backdrop, this article is intended as a primer for applied researchers working with LSAS as secondary analysts. Our exposition starts with a non-technical introduction to the key properties of skill assessment. We then review the three principal options that applied researchers have at their disposal to incorporate skill measures from LSAS in their secondary analyses: test scores, structural equation modeling (SEM), and plausible values (PVs). We discuss advantages and disadvantages of the three methods (i.e., test scores, SEM, and PVs) based on three criteria: *fallibility* (i.e., control for measurement error and unbiasedness), *usability* (i.e., ease of use in secondary analyses), and *immutability* (i.e., consistency across different analyses and analysts of test scores, PVs, or measurement model parameters in SEM). Our aim is to inform secondary analysts about the advantages and potential pitfalls of each option in order to help them make informed choices and understand potential limitations and biases that may ensue from using one option. The most important take-away message will be that using a single ability point estimate per person—that is, using test scores—is not the most appropriate option for research on skills. We conclude with practical recommendations.

From testing to test scores: three basic properties of skill assessment

Many innovations in psychometrics have sprung from the context of LSAS. To appreciate why increasingly sophisticated psychometric models are needed, it is critical to understand the properties of skill assessments and the challenges these entail: (1) the distinction between latent (unobserved) skill variables and their manifest (observed) indicators; (2) the concept of measurement error; and (3) the difference between measurement/population models and individual ability point estimates. Below, we briefly introduce these properties as they relate to LSAS.

Skills are latent variables

The cognitive or socio-emotional skills that LSAS seek to assess are *latent variables*. That is, these skills cannot be directly observed but only inferred from individuals' responses to a set of test items. This is typically expressed in path diagrams such as the one shown in Fig. 1. Here, observed responses on items 1 through K (i.e., Y_k) are used to estimate the latent skill, denoted θ, an idea first introduced by Spearman (1904) in his true score model.

Consider the example of literacy skills (i.e., the ability to understand, evaluate, and utilize written text; e.g., OECD, 2016). There is no way to directly measure an individual's literacy skills as one would measure their height or body weight. However, literacy can be made accessible to measurement if one conceived of literacy as a latent variable that manifests in individuals' ability to solve test items that were designed such that they require a certain level of literacy skills to be solved. Test takers' answers to test items are observed ("manifest") variables that reflect the unobserved ("latent") variable of interest, literacy skills.

Test items and test scores contain measurement error

Because a skill is a latent variable, any test designed to measure it will only imperfectly capture the individual's true ability θ_i. Individuals' responses to each test items will always reflect extraneous influences other than the skill that the test intends to measure, that is, they will contain *measurement error* ε_i. Possible sources of measurement error include, for example, random influences such as guessing, accidentally choosing the wrong answer despite knowing the correct one, or external disturbances during the testing session. Especially as test length increases, factors such as fatigue, loss of motivation, or practice effects, may also tarnish item responses. Measurement error is indeed an inextricable property of test items and, hence, of tests scores that, in their most basic form, are the sum across these test items (Lord & Novick, 2008). Consequently, any test score will only yield an *estimate* of that true ability, $\hat{\theta}_i$, and that estimate will contain measurement error: $\hat{\theta}_i = \theta_i + \varepsilon_i$.

Hypothetically, one might get closer to the true ability by administering a very large number of test items, or by repeatedly testing every individual many times. Akin to improving the "signal-to-noise ratio," increasing test length or testing on multiple occasions can add information about $\theta_i$while minimizing the influence of measurement error (i.e., can increase the reliability of the test). This applies only if the measurement errors of individual items ε_{ik} are random and independent of each other and

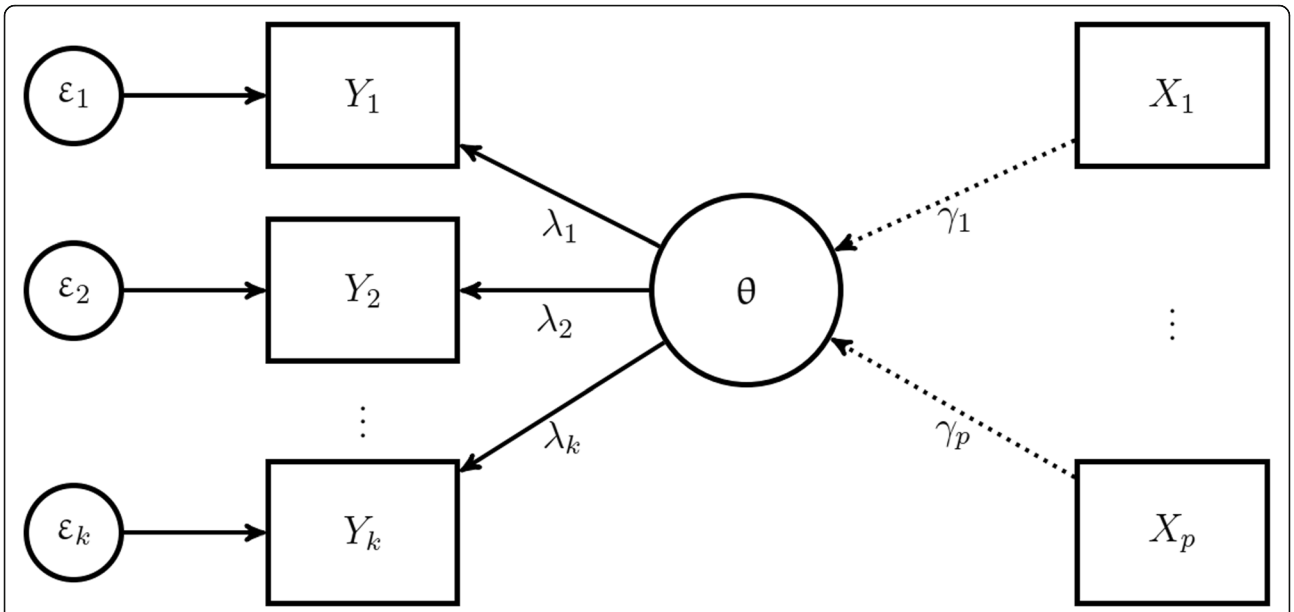

Fig. 1 A latent measurement model in CTT notation. The latent ability θ is measured by k test items (manifest indicators $Y_1 \ldots Y_k$). The different loadings $\lambda_1 \ldots \lambda_k$ and measurement error terms $\varepsilon_1 \ldots \varepsilon_k$ reflect different degrees to which each item reflects the latent ability. For substantive analyses and/or for estimating PVs, θ can be predicted by background variables ($X_1 \ldots X_p$). The latter model part is often referred to as the population model (e.g., Mislevy, 1991)

the additional items or measurement occasions are valid indicators of θ_i (for a brief exposition, see Niemi, Carmines, & McIver, 1986).[3] In real-world scenarios, however, resource constraints and concerns about respondent burden make it impossible to administer a large number of items, let alone administer them repeatedly.

Why should applied researchers care about measurement error? The answer is simple: If unaccounted for, measurement error can bias research findings. As has long been known (Fuller, 1987; Schofield, 2015; Spearman, 1904), classical measurement error (i.e., random error that is normally distributed and uncorrelated with the latent variable) acts like noise that blurs the signal. When error-laden skill measures are used as predictors in a regression, measurement error can substantially decrease the association between the skill and an outcome compared to its true size, a bias known as "attenuation bias" or "regression dilution" (Lord & Novick, 2008; Skrondal & Laake, 2001). A more specific variation of this problem occurs when researchers seek to control for confounders in a regression or to establish the incremental predictive validity of a skill over other predictors (or vice versa). Here, measurement error in a skill and/or one of the covariates can lead to overly optimistic conclusions about incremental validity (i.e., type I error; Westfall & Yarkoni, 2016) and phantom effects (i.e., biased compositional effects) in multi-level models (Pokropek, 2015; Televantou et al., 2015). In longitudinal studies, measurement error can reduce rank-order consistencies (stabilities) of skills. Moreover, although random measurement error does not bias estimates of population means, it does bias variances and hence also standard errors (i.e., precision).

In addition to classical measurement error, a variety of other biases other than attenuation can occur when using test scores (i.e., ability point estimates $\hat{\theta}_i$). These biases can lead to both over- and underestimations of regression coefficients, variances, and related statistics (Hyslop & Imbens, 2001; Jacob & Rothstein, 2016; Nimon, Zientek, & Henson, 2012). For example, if measurement error in a skill is not classical but correlates with the measurement error or true score of an outcome the skill is meant to predict, assumptions about the independence of error and true scores are violated. This may not only attenuate but also inflate the regression coefficients describing the skill–outcome relationship (for in-depth discussions, see Fuller, 1987; Hyslop & Imbens, 2001; Nimon et al., 2012; Stefanski, 2000). Moreover, as we will see later, different methods of computing test scores entail different forms of biases that can have different, and often hard-to-predict,

[3]The relation between test length and test reliability is expressed in the so-called Spearman–Brown formula, sometimes called "prophecy formula." This formula allows to predict how the reliability *Rel* of a test will change when extending the length l of the test (l = number of items) by a factor of k. It stipulates that $\mathrm{Rel}(k \times l) = k \times Rel / [1 + (k - 1) \times Rel]$.

consequences (e.g., some "shrinkage" estimators pull individuals' ability estimates towards the population mean, making extreme scores less extreme).

Measurement models and individual ability estimates are not the same

The goal of LSAS is *not* to provide ability estimates for individuals. Instead, their goal is to provide estimates of *population quantities* such as means and variances of the skill distribution or associations between skills and their predictors and/or outcomes. Compared to tests that are meant to inform decisions about individual test takers (e.g., college admission or employee recruitment tests), the tests in LSAS comprise far fewer items. Additionally, LSAS often use complex booklet designs in which each individual works only on a small subset of items: Test items from a large item pool are assigned to blocks, which are arranged in test booklets. After answering a set of common questions from the background questionnaire, each respondent works only on a randomly assigned booklet, that is, only on some of the item blocks (e.g., Braun & von Davier, 2017). Such "planned missingness" or "incomplete block" designs reduce respondent burden and cut costs for the data-producing organization (see Graham, Taylor, Olchowski, & Cumsille, 2006, for a general introduction).

Analyzing data from skill assessments, and especially from skill assessments employing complex test designs, requires specialized statistical models. These statistical models comprise a *measurement model* (or "latent variable model") linking individual responses to test items with the latent skill construct $\theta_i$and oftentimes also a *population model* stipulating the distribution of the latent skill (mostly the normal distribution) and its relations to the background variables from the background questionnaire (see Mislevy, 1991, pp. 180–181). Crucially, such models are primarily designed for estimating population parameters of *tests* (e.g., item difficulties or reliability) and the resulting skill distribution in the population—but not necessarily for providing individual ability estimates (i.e., test scores).

To better understand this point, let us briefly—and with some omission and simplification—review the psychometric theories that underlie skill assessments in LSAS: classical test theory (CTT) and item response theory (IRT), also known as probabilistic test theory (Lord & Novick, 2008; see Steyer, 2015, for an overview). Both theories provide measurement models that express the relation between a latent variable (e.g., a skill) and its indicators (e.g., test items) in different but closely related ways (Glöckner-Rist & Hoijtink, 2003; Raykov & Marcoulides, 2016). Most modern cognitive skill assessments in which test items have a binary or categorial response format (e.g., correct–incorrect) are based on IRT models. IRT is arguably better able to handle incomplete block designs, population models with background variables, and computer adaptive testing (e.g., Bauer & von Davier, 2017). CTT continues to be widely used and has an important place in scale construction, especially (but not exclusively) in the assessment of socio-emotional skills using polytomous (rating scale) formats. Many LSAS employ both CTT and IRT, albeit in different stages of the analysis. Without going into detail, both theories assume that responses to test items reflect the person's true ability—but only imperfectly and often to different degrees. The fundamental equations of these theories map the latent true ability to the observed answers to test items while highlighting that these manifest items are only imperfect (i.e., unreliable) indicators of the latent quantity. As we saw in Fig. 1, in CTT, a person's response to an assessment item (or subtest) Y_{ik} is modeled as a function of their true ability θ_i (often scaled by a factor loading λ_k that indicates how strongly the item reflects the θ_i) and a measurement error ε_{ik} that is orthogonal to true ability:[4]

$$Y_{ik} = \lambda_k \cdot \theta_i + \varepsilon_{ik} \quad (1)$$

Measurement error in CTT is defined as the difference between the observed response and the true ability, $\varepsilon_{ik} = Y_{ik} - \theta_i$. The main goal of CTT is to garner information about tests (not test takers). Of particular interest is a test's reliability in a sample, defined as the proportion of variance in the test that is due to variance in the true scores, $\mathrm{Rel}(Y) = \mathrm{Var}(\theta)/\mathrm{Var}(Y)$.

In IRT, each respondent's probability of answering an item correctly is modeled as a function of the latent ability θ_i, the difficulty of the item b_k, and often scaled by an item discrimination parameter a_k. For example, a model for a binary test item where wrong answers are coded 0 and correct answers coded 1 can be estimated as:

$$P(Y_{ik} = 1) = \exp(a_k \cdot (\theta_i - b_k))/[1 + \exp(a_k \cdot (\theta_i - b_k))] \quad (2)$$

Unlike in CTT, measurement error does not appear as a parameter in the IRT equation. Instead, it is implicit in the probabilistic (i.e., non-deterministic) relationship between the latent ability variable and its manifest indicators. Note that in IRT the term "measurement error" is often used to denote the standard error $SE(\hat{\theta}_i)$ of a re-

[4]CTT notation traditionally uses $\tau\tau_i$instead of $\theta_i$and uses the term "true score" instead of "true ability." Also note that the true score refers to the true score of a single test item or item parcel (indexed k, hence $\tau\tau_{ik}$) and is not necessarily identical to the latent ability θ_i(this is only true for the model of parallel tests; e.g., Steyer, 2015). We use $\theta\theta_i$instead of $\tau\tau_i$for simplicity and consistency across IRT and CTT.

Table 1 Evaluation of the three main methods of using skill measures from LSAS

Method	Variants and Examples	Fallibility	Usability	Immutability
Test scores	▪ Sum scores (weighted, unweighted) ▪ CTT factor scores (Thurstone, Bartlett, EAP) ▪ IRT ability estimates (WLE, MLE, EAP, MAP)	▪ ME not (fully) controlled (–) ▪ Biased standard errors of the latent variable in regressions (–) ▪ Biased variance estimates (e.g., underestimation for EAP, overestimation for WLE) (–) ▪ Factor score indeterminacy (–)	▪ Sum scores: Very easy to compute (+) ▪ CTT and IRT test scores, if user-generated: Computation requires knowledge of psychometric models but is fairly easy (+) ▪ Very easy to use in analysis (+)	Sum scores: ▪ Immutable across sub-samples, analyses, and analysts (+) CTT Factor scores/ IRT ability estimates: ▪ Immutable if estimates are included with LSAS data (+) ▪ Not immutable if estimates are user generated (–)
Structural equation modeling (SEM)	▪ Regular SEM ▪ IRT–SEM ▪ MESE	▪ ME controlled (+) ▪ Unbiased estimates of correlations, means, etc. of the latent variable (+) ▪ Measurement model sensitive to model (mis-)specification (–)	▪ Requires specialized statistical software (–) ▪ Requires additional psychometric expertise (–)	▪ Immutable if measurement model parameters are fixed (+) ▪ Not immutable with free measurement model parameters across sub-samples, analyses, and analysts (–)
Plausible Values (PV)		▪ ME controlled (+) ▪ Approximately unbiased estimates of correlations, means, etc. of the latent variable (+)	▪ User-generated PVs require statistical and programming and expertise (–) ▪ Using PVs in secondary analysis requires basic knowledge of multiple imputation methodology (–)	▪ Immutable if PVs are included with LSAS data (+) ▪ Not immutable if PVs are user generated (–)

Fallibility indicates whether the method accounts for measurement error and is unbiased. *Usability* denotes the ease of use in secondary analyses. *Immutability* is the property of test scores, PVs, or measurement model parameters *not* to change (i.e., remain the same) across different analysis setups and analysts
ME measurement error. See "Abbreviations" for all other abbreviations

spondent's test score, or more generally of the ability point estimate.

It may surprise readers to learn that CTT models can operate on an input matrix that solely contains the sample variances and covariances but not individual responses to test items. Many IRT models do operate on individual responses, yet the estimation of item parameters (the a_k and b_k) occurs prior to, and independent of, any estimation of person parameters (i.e., ability estimates).[5] Computing test scores thus involves a transition from a CTT or IRT measurement model for the population to a single point estimate $\hat{\theta}_i$ for an individual's unknown true ability θ_i. This step, called "ability estimation," is critical: Only a latent measurement model of a skill, but not a prediction of an individual's test score, is able to separate true ability from measurement error (e.g., McDonald, 2011). Moreover, some test scores such as unit-weighted sum scores ignore possible differences in item difficulties and discriminations in the measurement model (McNeish & Wolf, 2020; von Davier, Gonzalez, & Mislevy, 2009). Fortunately, as we will see, some approaches circumvent computing individual point estimates and the biases that can result.

Overview of the three main methods for using skill measures from LSAS

We now review the three principal methods that secondary analysts can use to incorporate skill measures from LSAS in their analysis: as test scores, through structural equation modeling (SEM), and in the form of plausible values (PVs). Table 1 provides an overview of the three methods.

After a general description of each method, we therefore evaluate it based on three criteria:[6] (1) fallibility, (2) usability, and (3) immutability. *Fallibility* describes whether the method accounts for measurement error; that is, whether it separates true ability from measurement error. Associations between fallible measures and predictors or outcomes of interest are subject to attenuation bias (i.e., lower than they truly are) and other forms of bias. Fallibility is the most important touchstone for comparing the methods, as it is most important for the unbiasedness of research results. *Usability* denotes the ease of use for secondary analysts in terms of the required statistical and data-analytical expertise. Usability is an important consideration because even methods that are less biased but too complex to implement tend to be less popular with secondary analysts and are generally prone to being used erroneously. Finally, *immutability* indicates whether (a) the individual ability point estimates (test scores), (b) PVs, or (c) parameters (e.g, the loadings λ_k) of a latent measurement model in SEM remain the same (i.e., unchanged) across different analysis setups (i.e., variables included in the

[6]These criteria correspond only loosely to statistical concepts. We use them as umbrella terms to summarize important information about the four methods.

analysis, subsamples used, statistical models and estimators employed) or not. If test scores, PVs, or parameters of the latent measurement models are *not* immutable, this may also lead to different statistical inferences about a parameter of interest (e.g., the relation of a skill to a predictor or outcome) and ultimately to different substantive conclusions. It is, of course, highly undesirable if different analysts arrive at different conclusions and policy implications merely because of variations in how they analyze the same data. Immutability is, thus, closely related to the replicability of research findings, for which it is an important prerequisite.

Regarding these criteria, it is important to distinguish between two scenarios. In the first scenario, secondary analysts re-use test scores (or PVs) that were computed by the data-producing agency and included in the data dissemination. This is the most common scenario and limits what can be done to remedy the issues with test scores that we will raise. In the second scenario, secondary analysts estimate their own custom set of test scores, SEM, or PVs. This increases flexibility on the side of analysts but requires specialized psychometric expertise. Even more fundamentally, it requires access to the item-level data (i.e., the data need to include variables that store information about test-takers' responses to individual test items), which is not always the case with LSAS.

For simplicity, we assume that the IRT or CTT measurement model for the skill in question is correctly specified. Further, we assume that there is no differential item functioning (DIF) or measurement non-invariance across subpopulations (i.e., the test functions alike in different subgroups). We also do not consider complications introduced by missing data that stems from respondents not reaching or refusing to answer some test items. Finally, we will not deal with issues such as scaling, scale anchoring, linking, or test score interpretation. These issues are far from trivial but are beyond the scope of our present paper. Fortunately, in modern LSAS, most of these issues are taken care of by the test developers and data producers at earlier stages. Thus, secondary analysts need not be overly concerned with them, although it is good practice to critically examine whether assumptions such as measurement invariance/differential item functioning have been tested and are met. We will refer the reader to specialized treatments of these issues in the following.

[6]These criteria correspond only loosely to statistical concepts. We use them as umbrella terms to summarize important information about the four methods.

Test scores

Definition and description

As explained in the previous section, test scores are point estimates of an individual's ability $\hat{\theta}_i$. They are the scores that would be reported back to test takers, for example in an admission or placement test, and used for diagnostic decisions. There are many different types of test scores that range from simple sum scores to more complex Bayesian techniques. All these techniques share the aim of maximizing validity by producing test scores that are as highly correlated with the underlying true ability as possible. Some of them are well suited for the purpose of individual diagnostics—however, all types of test scores share some fundamental limitations that make them less-than-optimal choices for secondary analysts of LSAS who are interested in population quantities (e.g., means or variances of skills, group differences in skills, or relations of the skill to instructional quality or other predictors). Below we briefly describe the most widely used types of test scores (all of which will typically correlate highly for a given assessment, but not all of which can be computed when complex assessment designs such as the aforementioned incomplete block designs are used).

Sum scores Sum scores are the simplest type of test score. They are what the term "test scores" traditionally referred to (Lord, 1980). Their abiding popularity stems from the fact that they are easy to compute and interpret. However, as we will see, this simplicity can be deceptive as it masks the shortcomings of sum scores. These shortcomings explain why sum scores are no longer widely used in LSAS. Beauducel and Leue (2013), McNeish and Wolf (2020), and von Davier (2010) provide excellent discussions of the limitations of sum scores.

Assuming we have three indicators (i.e., Y_{ik} with $k = 1, 2, 3$) to measure a skill, the sum score for person i is computed as

$$\text{Unweighted Sum Score}_i = Y_{i1} + Y_{i2} + Y_{i3} \tag{3}$$

Instead of the sum, one can also take the mean across items. Most commonly, sum or mean scores are unweighted (or unit-weighted) such that all items contribute equally to the resulting scale score. This is only valid if all test items reflect the target skill in equal measure and with the same amount of measurement error. These rather restrictive assumptions are foundational to the model of "parallel test" in CTT (see Steyer, 2015, for an introduction) and the one-parameter logistic model (1PL or "Rasch" model) in IRT (Andersen, 1977). In both models, all test items have the same factor loadings and error variance or item discriminations, respectively. This

assumption does not always hold in skill assessments, such that congeneric CTT models or (at least) two-parameter logistic IRT models (2PL or "Birnbaum"; Andersen, 1977; Birnbaum, 2008) are needed. According to these models, items can have different loadings or discriminations, which implies that they are not interchangeable and reflect the latent skill to varying degrees. In this case, unweighted sum scores are inappropriate because unit-weights do not align with the measurement model (Beauducel & Leue, 2013; McNeish & Wolf, 2020).

Researchers sometimes hope to remedy the problems of the sum scores by using weighted scores:

$$\text{Weighted Sum Score}_i = \lambda_1 \cdot Y_{i1} + \lambda_2 \cdot Y_{i2} + \lambda_3 \cdot Y_{i3} \quad (4)$$

The weights are typically taken from the loadings (λ_k) or item discriminations of the CTT or IRT model or from another dimension reduction technique such as principal component analysis (Joliffe & Morgan, 1992).

Sum scores are based exclusively on the available answers of respondents to test items. They cannot readily handle missing data (e.g., Mazza, Enders, & Ruehlman, 2015; see also Enders, 2010). This also implies that sum scores are ill-suited for complex test designs in LSAS (von Davier, 2010). These complex test designs involve planned missingness designs in which individuals answer different subsets of items. They also utilize Information from background questionnaire in order to improve the precision and efficiency with which $\hat{\theta}_i$ can be estimated. Items from the background questionnaire can also serve as "screening items" that govern which subset of items a respondent receives in booklet designs or computerized adaptive testing (CAT) designs. The skill data resulting from such designs cannot be readily summarized by a simple sum score.

CTT factor scores Factor scores are test scores based on factor-analytic CTT measurement models such as EFA and CFA. They are often used for computing test scores from socio-emotional (or "non-cognitive") skill assessments that use rating scale format. Factor score estimation methods account for both the factor loadings (i.e., the λ_k in Eq. 1 and Fig. 1) and the residual error variance information contained in the measurement model. There are several methods to compute factor scores, including the regression method, Bartlett's regression method, and expected a posteriori (EAP) estimator method (Beauducel, 2005; Devlieger, Mayer, & Rosseel, 2016; Fava & Velicer, 1992; Grice & Harris, 1998). For unidimensional measurement models, these methods result in different, albeit highly correlated, factor scores that are equally viable (Beauducel, 2007). One issue, especially for multi-dimensional measurement models (e.g., models with more than one factor) is factor score indeterminacy (Grice, 2001). Factor score indeterminacy means that an infinite number of factor scores can be computed from the same factor solution and all will be equally consistent with the model that produced the factor loadings. The higher the factor score indeterminacy, the higher the differences in the factor scores from different estimation methods. Factor score indeterminacy is lower when there are a large number of items and the items have strong factor loadings. To obtain consistent estimates of regression coefficients and their standard errors, Skrondal and Laake (2001) recommended using the Regression method to compute factor scores when the factor (e.g., a skill) is used as a predictor variable and Bartlett's method when it is used as an outcome variable in the subsequent regression analyses.

IRT ability estimates In modern LSAS, test scores are often computed from IRT models such as the 2PL, 3PL, or partial credit model (PCM). The two most widely used ability estimates from IRT models are likelihood-based methods such Warm's (1989) weighted likelihood estimate (WLE) and, once again, Bayesian methods such as the expected a posteriori (EAP) estimate. Whereas WLE depends only on the response pattern and the parameters of the measurement model, the EAP additionally depends on the prior distribution of θ. The EAP estimate is the mean of the posterior distribution of θ, which combines information about response patterns and model parameters with a prior distribution. Thus, unlike the WLE, EAP estimates can be computed with a prior distribution containing information from a background questionnaire (Laukaityte & Wiberg, 2017). Bayesian approaches such as EAP are inherently biased as they are shrinkage estimators, that is, the estimator pulls all test scores towards the mean of the prior distribution, thereby reducing their variance and making extreme scores less extreme. This bias is small when the prior distribution is appropriate and the reliability of the test is high (Tong & Kolen, 2010) but can be larger for test comprising only few items or when using incomplete block designs (e.g., Braun & von Davier, 2017). WLE and EAP estimates are widely used and tend to perform the best among other IRT-based ability estimates in terms of standard error of the regression coefficients in subsequent analyses.

Fallibility

As point estimates of individual ability, test scores turn the logic of latent measurement models that we showed in Eq. 1 upside down by predicting the latent ability from the observed items, rather than vice versa: $\hat{\theta}_i = \theta_i$

$+\varepsilon_i$ (e.g., McDonald, 2011). The fundamental problem that all types of test scores share is that the resulting point estimates contain measurement error, no matter how complex the model from which they were computed.

It is easiest to see this problem from the equations of the sum score (Eqs. 3 and 4). As per Eq. 1, the latent measurement model decomposes the answer to each item Y_{ik} into the (unobserved) true θ_i and a measurement error ε_{ik} in latent measurement models. By stark contrast, in the sum score equation, the items jointly determine the overall skill score. Measurement error in the items is not separated out from true ability but transferred to the sum score. Thus, building on CTT's logic, we can rewrite Eq. 4 as:

$$\begin{aligned}\text{Weighted Sum Score}_i = {} & \lambda_1 \cdot (\theta_i + \varepsilon_{i1}) + \lambda_2 \\ & \cdot (\theta_i + \varepsilon_{i2}) + \lambda_3 \\ & \cdot (\theta_i + \varepsilon_{i3})\end{aligned} \tag{5}$$

Because all individual test items Y_{ik} confound true ability and measurement error, the resulting sum score also contains measurement error. Only under the unrealistic assumption that no measurement error is present in the items would the sum score equal the (unobserved) ability θ_i.[7] Thus, sum scores are not infallible indicators of θ_i. Weighting the indicators as in the weighted sum score or principal component scores does not remedy this issue (Raykov, Marcoulides, & Li, 2017). Nor does using complex CTT or IRT models to compute test scores: Although the ability estimation process partly accounts for measurement error by considering the factor loadings or item discriminations in the measurement model, the resulting factor scores, WLEs, and EAPs are merely *realizations* of the random variable θ_i (Hardt, Hecht, Oud, & Voelkle, 2019; see also McDonald, 2011) and hence fallible point estimates that contain measurement error (Hoijtink & Boomsma, 1996). In other words, whereas latent CTT or IRT measurement models separate true ability from measurement error, all forms of test scores again compound them.

Moreover, depending on the ability estimation method used, test scores can contain additional biases. For example, because EAP is a "shrinkage estimator," EAP scores underestimate the population variance of the skill (Lu, Thomas, & Zumbo, 2005; Wu, 2005). The farther away an individual's score from the posterior mean (i.e., the mean after incorporating the prior distribution, which often contains information from background variables), the more it gets pulled towards the posterior mean. Contrariwise, WLE scores tend to have slightly lower conditional bias (i.e., the bias in the expected mean given θ) but higher standard deviation than EAP (Lu et al., 2005). Also, for both EAP and WLE, there are expected differences between individuals' ability estimates and their true ability scores, and these differences remain even in the case of large samples (Lu & Thomas, 2008).

As a consequence of the measurement error (and potentially other biases) contained in test scores, covariance-based statistics (e.g., correlations or regression coefficients) involving test scores can be biased. When the test scores are used to predict an outcome, the bias is often (but not invariably) attenuation or "regression dilution," such that the true size of associations between the skill and its predictors or outcomes are underestimated (Lord & Novick, 2008). Both EAP and WLE scores lead to deflated regression coefficients especially as the test length decreases (Braun & von Davier, 2017; Lu et al., 2005). The standard errors of regression coefficients are also biased since the variance estimate of the skills is biased.

Of note, different from variances and standard deviations, estimates of the skill's mean or mean differences across groups remain unbiased when using CTT factor scores. This is because CTT assumes that random measurement error has a mean of zero, which implies that the error is canceled out as one aggregates across a large number of items and individuals (Lord & Novick, 2008). Likewise, IRT ability estimates (EAP and WLE) both provide unbiased estimates of population means (Wu, 2005).

Thus, using test scores in secondary analyses can lead to biased estimates of population variances, regression coefficients when the skill is predictor (independent variable), standard errors—and hence potentially lead to erroneous conclusions in secondary analysis. It appears that the crucial difference between the latent CTT/IRT measurement models and the point estimates $\hat{\theta}_i$ computed from these models is not always clear to secondary analysts. The assumption that test scores derived from latent measurement models are somehow purified from measurement error is erroneous. Whether using factor scores, WLEs, or EAPs, no model-based ability estimates can remove the measurement error from $\hat{\theta}_i$ —and different methods can introduce different forms of additional bias.

Usability

Test scores are easy to understand conceptually, easy to compute, and easy to incorporate in secondary analysis. They can be treated much like any other variable (e.g., gender). Thus, the usability of test scores in the most common scenario—that is, secondary analysts working

[7] *Iff* the test items conform to a 1PL (Rasch) IRT model, then the sum score (at least) is a sufficient statistic for the latent ability θ_i.

with pre-computed test scores provided by the data-producing organization—is generally high.

Computing test scores is also straightforward in the case of sum scores (although, as noted, this no longer applies to complex test designs in modern LSAS that contain missing data by design). Computing factor scores and IRT ability estimates such as WLEs, and EAPs is somewhat more involved. However, provided basic familiarity with CTT or IRT, estimating measurement models and computing ability estimates from them is accessible through modern statistical software.

Immutability

In the scenario in which secondary analysts re-use test scores provided by data-producing organizations, test scores fulfill the immutability criterion: Test scores do not change depending on the covariates or subsamples used in the substantive analyses. They also do not depend on the analyst or analysis setup.

Conversely, when secondary analysts compute their own set of test scores from the original item-level data, these test scores are no longer immutable. This is because CTT factor scores and IRT ability estimates depend on the underlying measurement model, estimator, and the subset of respondents included in the estimation (Grice & Harris, 1998; Wainer & Thissen, 1987), and also on the population model if a population model is used (see Mislevy, 1991). Thus, substantive conclusions regarding the same research question using the same LSAS might differ between an analyst using test scores computed by the data-producing organization and another analyst using their own custom set of test scores.

Structural equation modeling

Definition and description

Provided access to the item-level data (i.e., variables that store individuals' answers to the single test items), structural equation modeling (SEM; Jöreskog, 1970; Jöreskog & Sörbom, 1979) offers a solution for measurement error in skill measures. Instead of computing fallible point estimates of ability from a measurement model, SEM combines the measurement model with a structural model. The *measurement* model (i.e., the relations of θ_i to the Y_{ik} in Fig. 1) represents the skill in question as a latent variable that is free from measurement error. The *structural* model relates this error-free latent variable to predictors, outcomes, or covariates through regression or correlation paths (e.g., the paths from the X_{ik} to θ_i in Fig. 1).

Thanks to dramatic advances over the last two decades, SEM has become an increasingly flexible and general approach (e.g., Bollen & Noble, 2011; Li, 2016). Routines for estimating SEM in modern statistical software can handle both continuous and categorical observed and latent variables, missing data, complex sampling designs, multiple groups, mediation and moderation, and many more scenarios relevant to LSAS. The measurement model part can either be based on a CTT or an IRT framework.[8] CTT and IRT measurement models are closely related (Glöckner-Rist & Hoijtink, 2003; Raykov & Marcoulides, 2016); both are usually estimated with maximum likelihood estimation (ML) or robust ML (MLR). Item factor analysis (IFA) using a weighted least squares estimator such as DWLS or WLSMV (designed to handle binary or ordered-categorical test items) can be seen as intermediate approach that bridges CTT and IRT (Glöckner-Rist & Hoijtink, 2003; Wirth & Edwards, 2007). For rating scales with five or more response options and data that are approximately normally distributed, different estimators lead to highly similar results (Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Hybrid approaches combine IRT with SEM (e.g., Lu et al., 2005). For instance, the mixed effects structural equations model (MESE; Junker, Schofield, & Taylor, 2012; see also Richardson & Gilks, 1993), extends the covariance-based general SEM framework for psychometric data (Bollen, 1989; Skrondal & Rabe-Hesketh, 2004). Here, the latent variable is defined via an IRT measurement model before the structural paths are added (Schofield, 2015). By using Bayesian priors, the MESE model allows users to condition the latent variable on covariates in the structural model to reflect extraneous influences on the latent ability.

Fallibility

Separating true ability from measurement error is the key motivation behind SEM and constitutes its main advantage over using point estimates of ability $\hat{\theta}_i$ such as sum scores in subsequent analyses. By *simultaneously* modeling the latent measurement model and the structural model, measurement error in the skill is separated from true ability (Jöreskog, 1969). Because SEM relates only the reliable portion of variance in the skill to other variables, relationships between the skill and predictors, outcomes, or correlates are unattenuated because they are purged from measurement error in the skill (although measurement error in the covariates, if unaccounted for, may still lead to attenuation bias).

Thus, SEM results in unbiased relationships between (latent) skills and other variables, avoiding biased results that would arise from using sum scores or model-based estimates (e.g., Fuller, 1987, 1995; Grice, 2001). SEM also largely avoids the additional biases on the population

[8]Note that latent variables may be *categorical* too, thereby suitable to model population heterogeneity and latent classes (e.g., Raykov, Marcoulides, & Chang, 2016).

variances and standard errors that using WLE or EAP test scores can entail. For the advantages of SEM to transpire, it is important that the measurement model be correctly specified (Mislevy, 1991; Rhemtulla, van Bork, & Borsboom, 2020), although it appears that regression coefficient estimates of the skills in the structural part are relatively robust to mis-specification of the measurement models or the conditioning population model of the underlying skill (Schofield, 2015). However, there is a risk that misspecification in the structural part of the model can affect the measurement model; just like adding or deleting covariates can influence the parameters of the measurement model—an issue that we turn to in the immutability section.

Usability

The disadvantage of SEMs compared to other methods of using skill measures from LSAS is their lower usability. Contrary to the other three methods reviewed here, secondary analysts must themselves implement SEM. This requires access to the item-level data (i.e., responses to individual test items must be included in the data). Each SEM is a unique analysis model tailored to a specific research question and cannot be disseminated with the LSAS data. Implementing SEM requires specialized statistical software and expertise that is not a routine part of the curriculum in all social and behavioral science disciplines. When IRT models (especially multidimensional IRT models) are used in the measurement part, only few software options are available and estimation can be computationally intensive (IFA with a DWLS or WLSMV estimator may provide a convenient solution here; Wirth & Edwards, 2007). Moreover, complex test designs (e.g., booklets or computer adaptive testing) may further complicate matters. For example, booklet designs or CAT can make it harder to specify and estimate a SEM due to the substantial amount of missing data. For these reasons, we expect that there will be only few occasions in which applied researchers employ SEM in secondary analysis of data from LSAS, despite the versatility of SEM and despite its advantages over test scores in terms of fallibility.

Immutability

Although the flexibility of SEM is an asset, it comes at the cost of violating immutability. If each analyst implements "their own" SEM, the parameters in the measurement model (such as the loadings λ_k relating the latent skill to its indicators; see Fig. 1) can change depending on a range of factors. These factors include not only the variables included in the measurement model but also those in the structural model (Anderson & Gerbing, 1988). Likewise, the (sub-)sample, missing data handling technique, and estimator used can result in different measurement model parameters and consequently different structural paths. If the measurement model parameters (say, factor loadings) change solely due to the presence (or absence) of predictors or outcomes in the structural model, the meaning of the latent variables changes and interpretational confounding occurs (Burt, 1973).[9]

A potential remedy for the mutability of measurement model parameters is using a two-step procedure in which the parameters of the measurement model are fixed after first estimating them in the absence of structural paths. The fixed parameters of the measurement model ensure that the latent variable is defined in a constant manner when structural paths are added in the second step (Anderson & Gerbing, 1988, 1992; see also, Bakk & Vermunt, 2016). In the context of IRT, this procedure is known as a "fixed IRT-SEM approach" (Lu et al., 2005, p. 271). Such a two-step procedure ensures that the latent measurement model is immutable across different researchers and specific research questions with specific sets of variables and paths.[10]

Plausible values

Definition and description

Originally developed in the context of NAEP (Mislevy, 1991), PV methodology is tailored to the needs of LSAS. Its aim is to provide unbiased estimates of population statistics such as means and variances of skills. Accessible introductions include , Bauer and von Davier, 2017, Lüdtke and Robitzsch (2017); in German), von Davier et al. (2009), and Wu (2005).

The basic idea of PVs is to treat ability estimation as a missing data problem and apply multiple imputation methodology (Little & Rubin, 2002; Rubin, 1987; for general introductions to multiple imputation, see Enders, 2010; Schafer & Graham, 2002; van Buuren, 2018). Instead of estimating a single test score $\hat{\theta}_i$ per respondent, multiple imputations of their unobserved true ability θ_i are generated. These imputations are called PVs

[10]We caution the reader about the lost variability of measurement model parameters once they are fixed. The uncertainty entailed in the measurement model is neglected when estimating structural paths. While this variability can be exploited to achieve higher parameter precision at large sample sizes, fixed IRT-SEM estimation still yields smaller finite sample bias than simultaneous IRT-SEM at smaller sample sizes (Lu et al., 2005).

[10]We caution the reader about the lost variability of measurement model parameters once they are fixed. The uncertainty entailed in the measurement model is neglected when estimating structural paths. While this variability can be exploited to achieve higher parameter precision at large sample sizes, fixed IRT-SEM estimation still yields smaller finite sample bias than simultaneous IRT-SEM at smaller sample sizes (Lu et al., 2005).

because they are educated guesses, based on a statistical model, of what a respondent's true ability might reasonably be. PVs are a special case of multiple imputations as the latent ability is completely missing. The observed information needed to impute the latent variables are their indicators (i.e., the test items Y_k in Fig. 1), the parameters of the measurement model λ_k linking these indicators to θ, and a population model containing a set of background characteristics such as gender, parental socio-economic status, or motivation variables (i.e., the X_k, also called "conditioning variables"). The latent measurement model from which PVs are computed can be any type of IRT or CTT model. Typically, five PVs per respondent are estimated and disseminated with the data, although more (e.g., 10–20) PVs may be preferable to obtain more precise estimates of standard errors (e.g., Laukaityte & Wiberg, 2017). The variation across PVs reflects the uncertainty about the respondent's true ability.

It is important to realize what PVs are *not*: They are not "test scores" in the traditional sense of Lord (1980); they are not point estimates of an individual's skills like CTT factor scores or IRT ability estimates. Also, PVs should not be confused with the true latent ability as conceived in CTT and IRT. Instead, PVs are intermediate quantities needed for the unbiased estimation of population quantities such as variances or regression coefficients (Bauer & von Davier, 2017).

More technically, PVs are repeated random draws from a posterior distribution that represents an individual's ability and the uncertainty about its true value. The posterior distribution $p(\theta_i, \boldsymbol{X_i}, \boldsymbol{Y_i})$ is "conditional" because it depends on the individual's responses to test items plus a (large) number of background variables contained in the latent regression model (Fig. 1). In formulaic notation, the posterior distribution from which PVs are drawn is

$$p(\theta_i, \boldsymbol{X_i}, \boldsymbol{Y_i}) \propto p(\boldsymbol{Y_i}|\theta_i) \cdot p(\theta_i|\, \boldsymbol{X_i}) \tag{6}$$

Here, $p(\boldsymbol{Y_i}|\, \theta_i)$ is the item response model that describes how the latent ability θ_i depends on the vector of item responses $\boldsymbol{Y_i} = (Y_{i1}, .., Y_{ik})$ of person i. Moreover, $p(\theta_i|\, \boldsymbol{X_i})$ is the population model that describes how the latent skill θ_i depends on a vector of background variables $\boldsymbol{X_i} = (X_{i1}, .., X_{ip})$:

$$p(\theta_i|\, \boldsymbol{X_i}) \sim N\left(\beta_0 + \boldsymbol{X_i}\, \beta_{\boldsymbol{p}}; \sigma^2_{\theta|X_i}\right) \tag{7}$$

The ability to incorporate background variables is a major advantage of PV methodology over traditional scoring methods such as sum scores or WLEs. Background variables often carry a great deal of information about an individual's likely standing on the skill scale, adding precision in estimating the PVs. Using this information for generating PVs allows LSAS to employ complex test designs (such as the booklet or "planned missingness" designs described earlier) that comprise far fewer test items than traditional designs (von Davier et al., 2009).

It may be instructive to note that there is a straightforward relationship between EAP test scores and PVs: As both are computed/drawn from the same posterior distribution, the EAP is the expected value across all PVs. This relationship makes it very evident that PVs adequately account for the uncertainty about the respondent's true ability whereas the EAP—as a single point estimate—does not.

From a practical perspective, incorporating PVs in secondary analysis involves the typical procedures for analyzing multiply imputed data: Each analysis (e.g., a regression model) is run once for each of the PVs. Parameter estimates are then pooled using "Rubin's rules" for means, regression coefficients, standard errors, and other quantities (Rubin, 1987). Of particular importance are the rules for pooling standard errors, which add uncertainty about the true ability. The uncertainty about the true ability is reflected in the variation across the different PVs per respondent and transferred to the variances and standard errors of parameter estimates. Therefore, the rules are necessary to obtain correct standard errors and p-values.

Fallibility

If used correctly—more on that later—PVs produce at least approximately unbiased estimates of population parameters such as means, variances, regression coefficients, and standard errors. Although it may not be immediately apparent, associations of the target ability with external variables such as predictors or outcomes of skills are corrected for random measurement error in the ability. Other biases incurred by computing test scores are also avoided. Estimates of population quantities based on PVs are (often much) closer to the true population value than those obtained with test scores (e.g., Braun & von Davier, 2017; Carstens & Hastedt, 2010; Laukaityte & Wiberg, 2017; von Davier et al., 2009; Wu, 2005).

As with SEM, the advantages of PVs only fully apply if the measurement model and population model used for generating the PV are correctly specified. Another precondition for unbiasedness that is not always met in LSAS is that if the data have a multilevel structure (e.g., students nested in schools), the PV-generating model must adequately reflect this structure (Laukaityte & Wiberg, 2017). Moreover, the PV-generating model must be at least as general as the analysis model. This precondition is known as the "congeniality" of the imputation model and analysis model (Enders, 2010; van

Buuren, 2018). Lest bias occur, the background model used for generating the PVs must include all variables (in transformed or untransformed form) that an analysis carried out by a secondary analyst includes (Bauer & von Davier, 2017). This pertains also to interactions between variables or higher-order (e.g., quadradic) terms. In practice, the conditioning model generating the PVs in LSAS typically includes a very large number of variables (indeed, all available) from the background questionnaire.[11] In this manner, it is possible to ensure that most conceivable analysis models will be (at least almost) congenial with the PV-generating model and to keep possible bias that stems from the omission of variables in the conditioning model to a minimum (von Davier et al., 2009). Thus, if PVs were generated based on a comprehensive conditioning model, secondary analysts need not be overly concerned with congeniality. Even with a comprehensive conditioning model, congeniality may be violated, however, if researchers introduce variables in the analysis model that were not part of the conditioning model or not even assessed in the background questionnaire. The latter can occur, for example, when secondary analysts match the data from LSAS with administrative records or geo-referenced data (e.g., regional unemployment rates).

Usability

It is possible for secondary data users to produce their own set of PVs, provided that they have access to the individual test items and possess the necessary analytical skills. Statistical software such as Mplus (Asparouhov & Muthén, 2010) or the R package TAM (Robitzsch, Kiefer, & Wu, 2020) have made PV estimation more accessible recently. Some LSAS such as the German NEPS (Scharl, Carstensen, & Gnambs, 2020) now provide dedicated tools for generating PVs based on a pre-specified measurement model and a custom population model in which secondary analysts can include the analysis variables required to achieve congeniality.

In most cases, PVs are provided by data-producing organizations. Although the lion's share of work is thus on the side of these organizations, it is fair to say that PV methodology does complicate matters for secondary analysts somewhat: Using PVs requires at least a basic understanding of multiple imputation methodology. Secondary analysts must run each analysis separately for each set of PVs and pool results using Rubin's rules.

It is important that secondary data analysts use PVs correctly, lest they lose the advantages of PV methodology for correct statistical inference. Two incorrect usages of PVs continue to be widespread (Jerrim et al., 2017; Laukaityte & Wiberg, 2017; Marchant, 2015; von Davier et al., 2009). The first incorrect usage is to use only one PV as if it were a point estimate, that is, a test score. Although this is the lesser sin and can produce unbiased estimates of population quantities (Wu, 2005), the uncertainty about each person's skill is lost and the variability information contained in the other PVs is neglected. The second mistake is to simply average across all PVs and use this average in subsequent analyses. Although the average across PVs produces a correct estimate of the ability's population mean, variances and standard errors will be biased downward as the uncertainty about the person's true ability is lost. Ignoring the uncertainty about the person's true ability may (but need not always; Marchant, 2015) lead to faulty inferences.

Fortunately, modern statistical software and modules make working with PVs less burdensome for secondary analysts. As working with PVs is the same as working with multiple imputations, programs such as Mplus, Stata, SPSS, or R all contain functions or packages that automate the process of working with PVs. For example, the Stata module REPEST (Avvisati & Keslair, 2020) facilitates analyses using PVs (and survey weights) from the PISA and PIAAC studies carried out by the OECD. Thus, using PVs has become straightforward at least for most standard analyses (e.g., multiple regression), and there is little reason to shy away from using PVs as a secondary analyst for usability reasons.

Immutability

As noted earlier, typically the data-producing organization generates a set of PVs intended to serve as broad a range of research questions as possible. In that case, immutability is assured because all secondary analysts will use the exact same set of PVs for their analyses.

If, by contrast, secondary analysts estimate their own PV (e.g., because they want to include additional conditioning variables to ensure congeniality), these PVs will differ depending on the specific type of measurement model chosen, the set of background variables included in the conditioning model, the subsample, and the imputation approach.[12] As a consequence, substantive results (e.g., relationship between the skill and some outcomes) obtained with different sets of PVs may also differ. Although user-generated PVs may help in

[12]To be precise, the specific PVs per respondent can change by chance alone even when re-running the exact same PV-generating model. This is because, as outlined earlier, PVs are random draws from the posterior distribution. This, however, will typically not affect pooled estimates from these two sets of PVs.

[12]To be precise, the specific PVs per respondent can change by chance alone even when re-running the exact same PV-generating model. This is because, as outlined earlier, PVs are random draws from the posterior distribution. This, however, will typically not affect pooled estimates from these two sets of PVs.

achieving congeniality between the PV-generating model and analysis model, they thus violate our immutability criterion. To prevent large discrepancies across analysts and analyses, LSAS could define a set of standard conditioning variables that secondary analysts should include in their PV-generating model.

Discussion

Secondary analysts working with data from LSAS often use test scores in much the same way as they use other analysis variables (e.g., gender, educational attainment). However, as our review highlighted, there are several problems with test scores that are not yet widely recognized by secondary analysts (Braun & von Davier, 2017; Jacob & Rothstein, 2016; Jerrim et al., 2017; von Davier et al., 2009): As point estimates of ability, test scores are not fully adequate for the task of statistical inference in LSAS. Test scores do *not* control for measurement error in the skill, leading to various biases in regression coefficients, standard errors, and other population statistics. Some types of test scores are also unable to handle modern LSAS's complex test designs and to incorporate information from background variables.

At this point, the reader may wonder how large and relevant the bias incurred by using test scores actually is. Simulation studies are best suited to answer this question. In simulation studies, data are simulated such that—unlike in real data—the true ability per simulated respondent is known, which allows quantifying various forms of bias. In one such study, Braun and von Davier (2017) studied the extent of attenuation bias that can occur in regression models in which a skill is an independent variable (i.e., predictor). The regression coefficient estimates based on five PVs were highly similar to the true population value, that is, unbiased. On the other hand, regression coefficients based on IRT ability estimates—EAP, WLE, and the simple maximum likelihood estimate (MLE)—were severely attenuated, with estimates 20–46% lower than the true population value of the regression coefficient (for details, see Table S1 in the Supplementary Online Material [SOM]). Moreover, EAP, WLE, and MLE (but less so PVs) produced overestimated regression coefficients for a covariate. These results were observed both in the case of congeniality and even non-congeniality between the generating model of PVs and the regression model used for substantive analysis.

Another simulation study by Wu (2005) showed that the population *mean* was correctly estimated not only by PVs but also by IRT ability estimates (WLE, EAP, and MLE) (see Table S2 in SOM). However, only PVs provided nearly correct estimates of population *variance,* whereas IRT ability estimates were biased for both 20-item and even more so for three-item tests. Similarly, von Davier et al. (2009) showed that the population means were predicted fairly accurately regardless of the number of items on the test and the scoring method used—EAP, EAP adjusted for group membership (EAP-MG), Warm's correction for MLE (WML), and five PVs (Table S3 in SOM). However, this was not the case for estimated population standard deviations, which only PVs were able to recover accurately. All other methods were biased, and, akin to the Wu (2005) study, bias increased as the number of items tested decreased.

In sum, these simulation studies highlight that population means of skills are unbiased when using test scores. However, the skills' variances, standard errors, and regression coefficients when using the skill as an independent variable will all be biased when using test scores, which may lead to erroneous statistical inferences (Braun & von Davier, 2017; Lu et al., 2005; Schofield, 2015; Wu, 2005). PVs perform well in all scenarios.

Practical recommendations

Based on our review, our recommendations for secondary analysts are clear: Whenever possible, secondary analysts should avoid using test scores in favor of methods that adequately account for measurement error in the target skill and preserve the uncertainty about the skill's true value per individual. In this regard, PV methodology lends itself as currently the best choice that is tailored to the needs of LSAS. If used correctly, PVs can prevent the various forms of bias in variances, regression coefficients, and their standard errors, as well as other population statistics that using test scores can entail. Moreover, using PVs can help avoid overly optimistic conclusions (i.e., type I error) in questions involving incremental predictive validity of some variable over a skill or vice versa (e.g., Braun & von Davier, 2017; see also Westfall & Yarkoni, 2016). The best option for secondary analysts in terms of fallibility, immutability, and usability is to use PVs provided by the data-producing organization and included in the data dissemination. If these PVs are based on an extensive background model, congeniality is typically a minor concern. If PVs are provided, researchers should follow the correct methodology (i.e., run the analyses on each PV and pool results following Rubin's rules) and refrain from averaging PVs or using only one PV.

In the increasingly rare cases in which only test scores (e.g., WLE or EAP scores) but no PVs are included in the data, secondary analysts should be wary of—and discuss transparently—the potential biases that can ensue from using test scores. Alternatively, provided that item-level information is available, researchers with advanced psychometric knowledge might decide to use SEM or estimate a set of PVs by themselves—a process that, for

example, NEPS now enables with a tool for PV generation (Scharl et al., 2020).

This leads us to our recommendations for data-producing organizations responsible for LSAS. In our view, data-producing organizations should provide a set of PVs for each skill measured in a LSAS, based on an extensive background model. The measurement model and population (background) model on which these PVs and/or test scores are based should be made transparent, and computer code should be provided such that secondary analysts can reproduce these PVs as well as modify the model as needed. For this purpose, the data should include item-level information (i.e., variables that capture responses to individual test items) needed for re-estimating the models on which PVs and/or test scores were based. Following these recommendations will widen the range of options available to secondary analysts, enabling them, for example, to estimate their own PVs and/or SEM, as opposed to having to rely on test scores or PVs from a potentially non-congenial background model. It will also contribute to greater research transparency (see also Jerrim et al., 2017).

Conclusion

There are good reasons for secondary analysts to gradually move away from using test scores—or at least to be mindful of the shortcomings of deceptively simple test scores in the context of LSAS, where the interest is in population quantities. As our review has shown, secondary analysts have two main options to avoid the potential biases that result from using test scores: (1) directly *modeling* measurement error in a SEM framework; or (2) *incorporating* measurement error in the analysis model through PV methodology. When using SEM, the modelled skills should invoke measurement models defined by the responsible data-producing organization (and accompanied by some recommended model parameters to foster immutability). Using PVs that are already included in the data (and that are ideally based on an extensive background model that ensures congeniality) seems to us the most sensible option under the criteria of fallibility, usability, and immutability.

In line with previous authors (e.g., Braun & von Davier, 2017; Lüdtke & Robitzsch, 2017; von Davier et al., 2009; Wu, 2005), we therefore recommend that secondary analysts—as well as organizations responsible for LSAS—fully embrace PV methodology. Although some time and effort are necessary to understand the basics of PVs, we believe the effort is worthwhile, as these methods will enable analysts to produce more rigorous and reliable research findings from LSAS to inform policy and practice. We hope that our primer provided a good starting point.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42409-020-00020-5.

Additional file 1 : **Table S1**. Results of simulation study conducted by Braun and von Davier (2017). **Table S2**. Results of simulation study conducted by Wu (2005). **Table S3**. Results of simulation study by von Davier et al. (2009).

Abbreviations

1PL: 1-parameter logistic model; 2PL: 2-parameter logistic model; CAT: Computerized adaptive testing; CFA: Confirmatory factor analysis; CTT: Classical test theory; DWLS: Diagonally weighted least squares; EFA: Exploratory factor analysis; IRT: Item Response theory; LSAS: Large-scale assessment surveys; MESE: Mixed effects structural equations; ML: Maximum likelihood; MLE: Maximum likelihood estimate; MLR: Robust maximum likelihood; NAEP: National Assessment of Educational Progress; NEPS: German National Educational Panel Study; PCM: Partial credit model; PIAAC: Programme for the International Assessment of Adult Competencies; PISA: Programme for International Student Assessment; PV: Plausible value; SEM: Structural equation model(ing); TIMMS: Trends in International Mathematics and Science Study; WLSMV: Weighted least squares (means-and-variance corrected)

Acknowledgements

The authors thank Maya Moritz for proofreading and copyediting. Open Access funding enabled and organized by Projekt DEAL.

Authors' contributions

All authors contributed to the ideation for this paper, wrote the first draft together, and revised it. The author order reflects the relative share of contributions. The author(s) read and approved the final manuscript.

Funding

This paper was supported by two grants to Clemens M. Lechner: a grant by the German Research Foundation (DFG), "Stability and Change in Adult Competencies" (Grant No. LE 4001/1-1); and a grant by the Germany Federal Ministry of Education and Research (BMBF), "Risk and protective factors for the development of low literacy and numeracy in German adults" (Grant No. W143700A).

Competing interests

The authors declare that they have no competing interests.

Received: 10 July 2020 Accepted: 28 December 2020
Published online: 19 January 2021

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*(1), 69–81. https://doi.org/10.1007/BF02293746.

Anderson, J., & Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach: Comment on Fornell and Yi. *Sociological Methods and Research*, *20*(3), 321–333. https://doi.org/10.1177/0049124192020003002.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411.

Asparouhov, T. & Muthén, B. (2010). Plausible values for latent variables using Mplus. Mplus Technical Report. http://statmodel.com/download/Plausible.pdf

Avvisati, F., & Keslair, F. (2020). REPEST: Stata module to run estimations with weighted replicate samples and plausible values. Retrieved from https://econpapers.repec.org/software/bocbocode/S457918.htm

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*, *23*(1), 20–31. https://doi.org/10.1080/10705511.2014.955104.

Beauducel, A. (2005). How to describe the difference between factors and corresponding factor-score estimates. *Methodology*, *1*(4), 143–158. https://doi.org/10.1027/1614-2241.1.4.143.

Beauducel, A. (2007). In spite of indeterminacy many common factor score estimates yield an identical reproduced covariance matrix. *Psychometrika*, *72*(3), 437–441. https://doi.org/10.1007/s11336-005-1467-5.

Beauducel, A., & Leue, A. (2013). Unit-weighted scales imply models that should be tested! *Practical Assessment, Research, and Evaluation*, *18*(1), Article 1. https://doi.org/10.7275/y3cg-xv71.

Bertoli-Barsotti, L. (2005). On the existence and uniqueness of JML estimates for the partial credit model. *Psychometrika*, *70*(3), 517–531. https://doi.org/10.1007/s11336-001-0917-0.

Birnbaum, A. L. (2008). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Information Age Publishing (Original work published in 1968).

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179.

Bollen, K. A., & Noble, M. D. (2011). Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15639–15646. https://doi.org/10.1073/pnas.1010661108.

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assessments in Education*, *5*(1), 5–17. https://doi.org/10.1186/s40536-017-0050-x.

Burt, R. S. (1973). Confirmatory factor analytic structures and the theory construction process. *Sociological Methods and Research*, *2*(2), 131–190. https://doi.org/10.1177/004912417300200201.

Carstens, R., & Hastedt, D. (2010). The effect of not using plausible values when they should be: An illustration using TIMSS 2007 grade 8 mathematics data. In *4th IEA International Research Conference (IRC-2010), Gothenburg, Sweden* https://www.iea.nl/sites/default/files/2019-04/IRC2010_Carstens_Hastedt.pdf.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770. https://doi.org/10.1177/0013164415607618.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Fava, J. L., & Velicer, W. F. (1992). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research*, *27*(3), 301–322. https://doi.org/10.1207/s15327906mbr2703_1.

Fuller, W. A. (1987). *Measurement error models*. Hoboken: Wiley.

Fuller, W. A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, *63*(2), 121–141. https://doi.org/10.2307/1403606.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*(4), 544–565. https://doi.org/10.1207/S15328007SEM1004_4.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. https://doi.org/10.1037/1082-989x.6.4.430.

Grice, J. W., & Harris, R. J. (1998). A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research*, *33*(2), 221–247. https://doi.org/10.1207/s15327906mbr3302_2.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*, (6th ed.,). Pearson Prentice Hall.

Hardt, K., Hecht, M., Oud, J. H., & Voelkle, M. C. (2019). Where have the persons gone?–An illustration of individual score methods in autoregressive panel models. *Structural Equation Modeling*, *26*(2), 310–323. https://doi.org/10.1080/10705511.2018.1517355.

Hoijtink, H., & Boomsma, A. (1996). Statistical inference based on latent ability estimates. *Psychometrika*, *61*(2), 313–330. https://doi.org/10.1007/BF02294342.

Hyslop, D. R., & Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, *19*(4), 475–481. https://doi.org/10.1198/07350010152596727.

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives*, *30*(3), 85–108. https://doi.org/10.1257/jep.30.3.85.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review*, *61*, 51–58. https://doi.org/10.1016/j.econedurev.2017.09.007.

Joliffe, I. T., & Morgan, B. J. T. (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research*, *1*(1), 69–95. https://doi.org/10.1177/096228029200100105.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. https://doi.org/10.1007/BF02289343.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239–251. https://doi.org/10.1093/biomet/57.2.239.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639. https://doi.org/10.2307/2285946.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. University Press of America.

Junker, B. W., Schofield, L. S., & Taylor, L. (2012). The use of cognitive ability measures as explanatory variables in regression analysis. *IZA Journal of Labor Economics*, *1*(1), 4. https://doi.org/10.1186/2193-8997-1-4.

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in statistics - Theory and Methods*, *46*(22), 11341–11357. https://doi.org/10.1080/03610926.2016.1267764.

Li, C. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Routledge. https://doi.org/10.4324/9780203056615.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age Publishing (Original work published in 1968).

Lu, I. R., & Thomas, D. R. (2008). Avoiding and correcting bias in score-based latent variable regression with discrete manifest items. *Structural Equation Modeling*, *15*(3), 462–490. https://doi.org/10.1080/10705510802154323.

Lu, I. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling*, *12*(2), 263–277. https://doi.org/10.1207/s15328007sem1202_5.

Lüdtke, O., & Robitzsch, A. (2017). Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung. *Diagnostica*, *63*(3), 193–205. https://doi.org/10.1026/0012-1924/a000175.

Marchant, G. J. (2015). How plausible is using averaged NAEP values to examine student achievement? *Comprehensive Psychology*, *4*, Article 1. https://doi.org/10.2466/03.CP.4.1.

Martin, M. O., Mullis, I. V., & Hooper, M. (2016). Methods and procedures in TIMSS 2015. https://timssandpirls.bc.edu/publications/timss/2015-methods.html

Mazza, G. L., Enders, C. K., & Ruehlman, L. S. (2015). Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research*, *50*(5), 504–519. https://doi.org/10.1080/00273171.2015.1068157.

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika*, *76*(4), 511–536. https://doi.org/10.1007/s11336-011-9223-7.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-020-01398-0.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457.

Niemi, R. G., Carmines, E. G., & McIver, J. P. (1986). The impact of scale length on reliability and validity: A clarification of some misconceptions. *Quality and Quantity*, *20*, 371–376. https://doi.org/10.1007/BF00123086.

Nimon, K., Zientek, L. R., & Henson, R. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology*, *3*(102), 1–13. https://doi.org/10.3389/fpsyg.2012.00102.

OECD (2016). *The survey of adult skills: Reader's companion*, (2nd ed.,). OECD Publishing. https://doi.org/10.1787/9789264258075-en.

Oranje, A., & Ye, L. (2014). Population model size, bias and variance in educational survey assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, (pp. 203–228). CRC Press. https://doi.org/10.1201/b16061

Pokropek, A. (2015). Phantom effects in multilevel compositional analysis: Problems and solutions. *Sociological Methods & Research*, *44*(4), 677–705. https://doi.org/10.1177/0049124114553801.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, *76*(2), 325–338. https://doi.org/10.1177/0013164415576958.

Raykov, T., Marcoulides, G. A., & Chang, C. (2016). Examining population heterogeneity in finite mixture settings using latent variable modeling.

Structural Equation Modeling, *23*(5), 726–730. https://doi.org/10.1080/10705511.2015.1103193.

Raykov, T., Marcoulides, G. A., & Li, T. (2017). On the fallibility of principal components in research. *Educational and Psychological Measurement*, *77*(1), 165–178. https://doi.org/10.1177/0013164416629714.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220.

Richardson, S., & Gilks, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, *12*(18), 1703–1722. https://doi.org/10.1002/sim.4780121806.

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test analysis modules. R package version 3*, (pp. 5–19) http://cran.ma.imperial.ac.uk/web/packages/TAM/TAM.pdf.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. http://dx.doi.org/10.1002/9780470316696

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147.

Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6 (NEPS Survey Paper No. 71)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.lifbi.de/Portals/13/NEPS%20Survey%20Papers/NEPS_Survey-Paper_LXXI.pdf

Schofield, L. S. (2015). Correcting for measurement error in latent variables in used as predictors. *The Annals of Applied Statistics*, *9*(4), 2133–2152. https://doi.org/10.1214/15-AOAS877.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. https://doi.org/10.1007/BF02296196.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modelling: Multilevel, longitudinal and structural equation models*. Chapman & Hall. https://doi.org/10.1201/9780203489437

Spearman (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159.

Stefanski, L. A. (2000). Measurement error models. *Journal of the American Statistical Association*, *95*(452), 1353–1358. https://doi.org/10.2307/2669787.

Steyer, R. (2015). Classical (psychometric) test theory. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences*, (2nd ed., pp. 785–791). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.44006-7.

Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L. E. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement*, *26*(1), 75–101. https://doi.org/10.1080/09243453.2013.871302.

Tong, Y., & Kolen, M. J. (2010). IRT proficiency estimators and their impact. In *Annual conference of the National Council of Measurement in Education, Denver, CO* http://images.pearsonassessments.com/images/tmrs/tmrs_rg/9_IRT_Estimator_Scoring_42210.pdf.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*, (2nd ed.,). CRC/Chapman & Hall. https://doi.org/10.1201/9780429492259

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. https://doi.org/10.2307/25792024.

von Davier, M. (2010). Why sum scores may not tell us all about test takers. *Newborn and Infant Nursing Reviews*, *10*(1), 27–36. https://doi.org/10.1053/j.nainr.2009.12.011.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*, *2*, 9–36 https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*(4), 339–368. https://doi.org/10.3102/10769986012004339.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS one*, *11*(3), Article e0152719. https://doi.org/10.1371/journal.pone.0152719.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005.

Zhu, Y., Steele, F., & Moustaki, I. (2017). A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome. *Structural Equation Modeling*, *24*(5), 643–656. https://doi.org/10.1080/10705511.2017.1324310.

Publisher's Note